Review of WHO Table B1 (VanTreeck, et al., 2023) by The Institute for Research & Evaluation (IR&E or IRE)

Table B1: Discrepancy Summary [from VanTreeck, et al., 2023[a]]				IRE Analysis of "Discrepancies" in Table B1
Study	**Measurement**	**Study Replication Findings** **[as interpreted by VanTreeck, et al., 2023[b]]**	**IR&E Report Findings** **[as interpreted by VanTreeck, et al., 2023[b]]**	**IRE Comments and Color-Coded Conclusions** **VIOLET=A mistaken interpretation of study results by WHO** **YELLOW=A misunderstanding of IRE Table 7 by WHO** **BLUE=A reasonable decision by IRE on study interpretation** **GREEN=A confirmed discrepancy between IRE and study results**
Aderibigbe, 2008	Any condom use	Evidence of programme potential at 3 months. Table 5 indicates that pre-intervention data demonstrated no significant difference amongst groups ($p = 0.8761$). Study participants who used condoms increased by 36.5% and control participants who used condoms increased by 1.8%. The post-intervention data were significantly different ($p = 0.0003$).	Report indicated that this was measured at 3 months, but did not colour code it to indicate the impact.	This is a misunderstanding by the WHO reviewers (VanTreeck, et al) of the IRE data table, Table 7. The "3 months" entered in the table cell does indicate a significant 3-month impact, not just that a 3-month follow-up effect was measured. Where the outcome was measured but the impact was not significant, the table says "NS," as indicated in the Key at the bottom of Table 7. This was done consistently throughout the IRE data tables. The blue color-coding to indicate "program potential" was an interpretation imposed by IRE, not the study author (Aderibigbe), and failure to color-code this finding as such is an oversight on the order of a "typo," not a discrepancy between the IRE report and the study findings reported by Aderibigbe.
Aderibigbe, 2008	Frequent/recent sex	Evidence of programme potential at 3 months. Table 3 reports the number of sexual partners in the past 3 months (recent sex). The number of participants who reported 0 partners increased from 8 to 30 in the study group and decreased from 11 to 8 in the control group. The pre-intervention values were not significant (no difference among groups), but the post-intervention values were significant ($p = 0.0002$). This demonstrates the significant positive effect of the intervention on recent sex. Data were measured for Frequency of Sex (Table 4), but it was not significant.	Reported as NS.	"Number of Sex Partners in the Past 3 months" is by definition a measure of "Number of Sex Partners," not "Recent Sex." This is where we correctly recorded it. Because an increase occurred in those reporting 0 partners in the past 3 months, that change could have been recorded as a change in Recent Sex, but where there was also a reduction in those reporting multiple partners, it seemed more appropriate to register these results as an overall reduction in Number of Sex Partners, as shown in IRE's Table 7. Frequency of Sex was also measured, and this nonsignificant outcome was recorded in the IRE column labeled Frequent/Recent Sex. This was not a discrepancy between the IRE data table and the findings reported by Aderibigbe, merely a reasonable decision on how to record these findings within the categories and format that IRE had set up to use for its report.
Aderibigbe, 2008	# sex partners	Evidence of programme potential at 3 months. Table 3 reports the number of sexual partners in the past 3 months. The % of study participants who reported 2-4 partners decreased by 20.3% while the % of control participants increased by 1.7% (pre intervention showed no significant difference, post intervention $p = 0.0431$). The % of study particpants who reported >4 partners increased by 0.2% and the % of control participants increased by 3.7% (pre intervention $p = 0.0049$, post intervention $p = 0.0011$).	Reported that there were data at 3 months, but did not colour code it to indicate the impact.	This is a misunderstanding of the IRE data table, not a discrepancy between the study findings and the IRE report. See the first note in yellow in this column, above.
Agha, 2004	Frequent/recent sex	NS. Table 6, "Has regular partner" data are not significant, $p > 0.05$.	Reported as evidence of programme potential at 6 months.	Table 6 reports a significant effect for "Had casual partner past 3 months" (AOR=.33, 95% CI= .14-.78) at the 6-month follow-up. This is a measure of *recent sexual activity* and IRE reported this as Frequent/Recent Sex at 6 months.
Ajuwon, 2007	Any condom use	NS. Tables 3 and 4 report the condom use baseline and post-intervention measurements. Respectively, $p = 0.96$ and $p = 0.51$.	Reported as evidence of programme potential.	The Chi-square p-values reported in Tables 3 & 4 are not a test for pre-post change, in other words, for a program effect. The study text ("Effects of the Intervention," p.54) reported a significant pre-post increase in Condom Use for one of the 3

				intervention types (E3, p<.05). The Control Group did not experience a similar significant increase, indicating a program effect for Intervention Group E3. This, and a significant increase in condom use for Group E2 (p<.05), are reported as positive program effects in the study abstract.
Ajuwon, 2007 [2 data points]	Sexual Initiation	Positive and negative results at 9 months. Teacher instruction (E1) and mixed intervention (E3) had a lesser increase than the control while the peer led intervention (E2) did not. Proportion that had ever had sex: E1 0.4% increase, E2 9.8% increase, E3 4.6% increase; Control 5.7% increase. *P* values for baseline and post-intervention: 0.022 and 0.0034, respectively.	Reported as NM.	1. The text of the study reported this outcome as "having sexual experience in the 3 months preceding the survey" which would be considered a measure of "Recent Sex" not "Sexual Initiation." IRE used this definition and marked Sexual Initiation, NM (not measured). *This was not a research error by IRE*. 2. The Chi-square p-values reported in Tables 3 & 4 are not a test for pre-post change, in other words, for a program effect. They are cited as such in error by VanTreeck. The text *("Effects of the Intervention")* reported a significant pre-post increase in this measure for one of the Intervention types (E2) and also for the Control group, with no reductions in the other groups. Since the Control group also increased significantly, and there was no statistical comparison of the amount of increase for E2 vs. Control, we did not designate it as evidence of a negative program effect but as a lack of evidence of positive effect, i.e., NS for Frequent/Recent Sex. The study authors (Ajuwon, et al.) also did not report this as a negative program effect.
Borgia, 2005	Negative Effect	Negative effect. Yes, condom use and unprotected sex.	Reported as no.	The pre-to-post increases in sexual risk behaviors in Figure 1, though significant, cannot be attributed to a negative program effect. The reported changes merely indicate that the sexual risk behaviors increased significantly over time, which is not unusual in a teen population. Without a control group, it is unknown whether these increases might have been even greater without the program's influence, which would mean a positive program effect. There was no control group so the measure of a program effect was to test differences between the programs in these outcomes, and no differences were detected, per the first row of results in Tables 4 & 5. This claimed IRE discrepancy is actually a misinterpretation of the study's results by VanTreeck, et al., not an error in the IRE report.
Borgia, 2005	CCU	Negative effect. Decrease in always using condoms: Teacher led from 29.0 to 22.3%, Peer led from 36.8 to 27.6% ($p < 0.05$).	Reported as NM.	See above note (Borgia, 2005) regarding the error by VanTreeck, et al., in interpreting this study's results. Also, the analysis that measured for a program effect on condom use (Table 5) was for the outcome of Frequency of Condom Use" and did not specify Consistent Condom use (CCU) meaning "always" use in this study. Thus, *CCU was not measured in the outcome analysis of program impact.*
Borgia, 2005	Any condom use	Negative effect. Figure 1 shows the changes in condom use for both the teacher and peer interventions. While both interventions saw more participants responding to the condom use survey question with "sometimes use a condom", this was due to a decrease in "always use a condom". This data were significant ($p < 0.05$).	Reported as NS.	Again, this is an error by the WHO reviewers. Figure 1 does not measure a program effect but merely the change over time in the outcomes of interest for the 2 programs. There was no control group and no difference between programs (see Table 5 and the first IRE note on Borgia, 2005, above). The study authors reported no significant program effect, positive or negative, on condom use (see study abstract). "Neither of the interventions induced changes in sexual behavior." (Abstract)
Borgia, 2005	# sex partners	Negative effect. Teacher intervention had significant increase in respondents who reported having more than one partner	Reported as NS.	See above notes (Borgia, 2005) re. the incorrect interpretation of Figure 1 by VanTreeck, et al. The study authors reported no significant effect, positive or negative, on # sex partners (see study abstract).

		(Figure 1, $p < 0.01$). Peer intervention was not significant.		
Dente, 2005	Frequent/recent sex	Evidence of programme potential. Study states: "students of group 1 and group 2 were less likely to have had intercourse with casual partners than those of group 3 ($p = 0.006$)". Groups 1 and 2 are intervention groups; group 3 is the control group.	Reported as NM.	The study actually states (p.4), "among those who were recently sexually active, students of group 1 and group 2 were less likely to have had intercourse with casual partners than those of group 3 ($p = 0.006$)". Thus, "recent sex" was a given and not measured as an outcome itself—it does not appear as an outcome on Table 2, thus, it was Not Measured (NM).
Dente, 2005	Sexual Initiation	Positive effect. Table 2 presents data of [those] who have ever had sex in their lifetime. Voluntary Counselling and Testing (VCT) & School Health Education (SHE): 31.4%, SHE: 57.1%, Control: 62.9%. $p < 0.001$.	Reported as NS.	IRE concurs with VanTreeck, et al., that the IRE data table should show a post-program effect. However, IRE does not agree that this is evidence of program effectiveness, since the outcome measures were taken at the End of the Program; there was no evidence of a sustained post-program effect. (p.2 says it was "a post-test only…study")
Diaz, 2005; Salvador	Unprotected sex	Evidence of programme potential. Table 6 reports an AOR (study:control) of 2.82 (95% CI: [1.45–5.49]) for current use of modern contraceptive methods. Note, this was at no specified time period after the intervention because these groups were identified from actual school settings.	Reported as NM.	The measure "Use of modern contraceptive methods" usually includes the use of the birth control pill and/or LARCs[c] and may or may not include condom use. (It does not appear to in this study.) This measure is not an indicator of Unprotected Sex, because it is unknown whether condom use is occurring, and only condoms provide protection from STIs.
Doyle, 2010	Any condom use	Evidence of programme potential at 3 years in subgroup. Table 3 reports a higher condom use with non-regular partner among females in the intervention compared to those in the control (aPR 1.34 $p < 0.05$).	Reported as NS.	Where this study employed multiple ways of measuring the same basic outcome of condom use, IRE chose to use the more robust measure as the more valid one. "Condom use with non-regular partner" is not a measure of overall condom use, as is "condom use at last sex in past 12 months," which is the robust measure IRE used. It was not statistically significant for males or females, and this is what IRE reported in its data table. Rather than being an error by IRE, this was a reasonable decision on interpretation of the study outcomes.
Doyle, 2010	# sex partners	Evidence of programme potential at 3 years in subgroup. The intervention was associated with a reduction in the proportion of males reporting more than four sexual partners in their lifetime (aPR 0.87, 95%CI 0.78–0.97)	Reported as NS.	Where this study employed multiple ways of measuring the outcome of Number of Sex Partners, IRE made a judgment call, prioritizing the more recent measures as the more valid indicators. There were 4 measures related to Number of Sex Partners in this study. Although Lifetime Partners was reduced for males, there was no reduction in the other 3 measures ("in past 4 weeks," "in past 12 months," and "in the same time period in past 12 months"), all of which were much more recent measures. IRE concluded that the outcome, Number of Sex Partners was not significantly impacted for males. This was a reasonable decision, not an error by IRE.
Duflo, 2006	Pregnancy	NS. Table 6 reports childbearing rates. Data for "has started childbearing" is not significantly different between the teacher training and reducing the cost of education programmes.	Reported as NM.	"Childbearing" was not considered a measure of "Pregnancy" or "Ever been pregnant" by this study. On p. 15, the study authors state: "Of course childbearing is an imperfect proxy for risky sexual behavior…Abortion…could…reduce the correlation between the incidence of childbearing and the risk of HIV infection." In other words, the measure of childbearing doesn't take into account abortion, which may increase if pregnancy rates go up, leaving a childbearing rate that appears stable while actual pregnancy rates have increased. Thus, IRE reported Pregnancy as Not Measured.
Duflo, 2006	# sex partners	NM	Reported as NS.	Table 5 reports on "Had more than one sex partner" and it was NS. A reduction in the number of students who have had more

				than one sex partner is the same as measuring a reduction in the Number of Partners for the treatment group.
Duflo, 2015	Sexual Initiation	NS. Table 6 reports long-run impacts on participants' responses to "Ever had sex". Data were only significant at 10% level.	Reported as NM.	Duflo's Table 6 does not have data on the Critical Thinking (CT) program, which was the only program component with content on condom use and thus the only CSE study arm. Since IRE's focus was CSE, the results for CT were the only outcomes relevant to its report. IRE's data table specifies "Critical Thinking" as the program of interest. Table, 5, which does include the CT study arm, does not report on Sexual Initiation, so IRE coded it, correctly, as Not Measured (NM) for the Critical Thinking program. This is not an IRE error.
Duflo, 2015	Any condom use	NS. Table 5 presents data for "Used condom at last sex" but $p > 0.05$.	Reported as NM.	IRE concurs.
Dupas, 2011	Pregnancy	NS. Table 3 reports the childbearing probability. The RR data are not significant at a 95% CI ($p = 0.10$). The TT data do not have any significant childbearing data.	Reported as positive and negative effect.	Dupas' Table 6 shows negative effects on Sexual Initiation, Number of Sex Partners, and Paid Sex, all at p<.05. IRE's Table 7 labels these negative effects correctly, in Column 4. It does not report negative program effects on the outcome of Pregnancy. This is a misunderstanding of the color-coding on the IRE data table. The color green was used to indicate "Evidence of Program Effectiveness: A significant effect on a key protective indicator, at least 12 months post-program, on the intended target population (not just a subgroup), *without other negative effects*." Where a program had a positive 12-month effect on a key outcome, but a negative impact on another outcome, the positive 12-month effect was color-coded grey rather than green, to indicate that it was not "Evidence of Program Effectiveness" because of the occurrence of negative effects on other program outcomes. The grey color did not mean there were both positive and negative effects on the same outcome, but that *the program* produced both positive and negative effects across the targeted outcomes. This was noted in the color key at the bottom of Table 7, but could have been stated more clearly by IRE. Regarding a positive effect on Pregnancy: Table 3 in Dupas, 2011, reports a decrease in childbearing probability (considered by the study as a proxy for pregnancy) for unmarried girls as significant at p<.05. *Unmarried* girls was the group for which the outcome of Pregnancy was of interest to the program; for this reason it was entered in the IRE data table correctly as a positive program effect on Pregnancy.
Dupas, 2011	Frequent/recent sex	NS. Table 6 presents information for "Currently has regular partner". I would say that this qualifies as frequent/recent sex as this context is referring to regular sexual partner.	Reported as NM.	"Regular sex partner" does not measure how recently or how frequently sexual activity occurs, as these measures are usually construed. Thus, these outcomes were reported by IRE as Not Measured (NM). This is not an error by IRE but simply a difference in measurement interpretation.
Fitzgerald, 1999	Any condom use	Evidence of programme potential at 6 months. Table 5 reports data for condom use (used condom). The difference in condom use between intervention and control youth at follow-up (78% vs. 64%) was very similar to that at baseline (78% vs. 67%), although the difference at follow up reached statistical significant ($p < .05$).	Reported as negative effect.	Table 5 shows Condom Use for Intervention boys at baseline was significantly higher than Control boys (87% vs.68%, p<.05) but decreased for Int. boys at Wave 2 more than for Control boys (79% vs. 67%--Control boys essentially stayed the same) such that it was no longer significantly different, a statistical indicator of negative program effect on the boys. This is not an IRE error but rather a reasonable decision on interpretation of study results. The overall effect was due to an increase by girls.

Fitzgerald, 1999	Dual benefit (12 mo.)	NS. Due to the discrepancy for Fitzgerald CCU, this is also a discrepancy. Sexual initiation was measured but was NS.	Reported as NM.	The follow-up period was 6 months (p.55) therefore a 12-month Dual Benefit was not measured (NM).
Fitzgerald, 1999	CCU	NS. Among intervention youth, the percentage of youth reporting frequent use of condoms ("always"/"usually") increased 16% (from 61% at baseline to 77% at follow-up) compared to only a 4% increase (64% to 68%) among control youth (p = not significant).	Reported as NM.	"Always/usually" using a condom is a measure of "frequent use of condoms," according to the study authors, and is not a measure of Consistent Condom Use (CCU), which would mean always using a condom or using a condom every time. This is a commonly understood distinction in sex education research and not an error by IRE.
James, 2006 [2 data points]	Sexual Initiation	NM. The study does not discuss data on sexual initiation. Additionally, the original report marks this as programme potential at 4 months, but there is only f/u at 6 and 10 months.	Reported as measured at 4 months; coded as evidence of programme potential.	1. On p. 286, the study says participants were asked to report "whether they were ever sexually active (0 = no, 1 = yes)." This is a measure of Ever Had Sex or Sexual Initiation. They were also asked whether they had had recent sex. Because the outcome reported was "sexual activity," without distinguishing between the 2 measures, we coded it as an effect for both. 2. The 6-month post-test measure was taken "directly after the intervention" (p.287), indicating it was a 6-month post-baseline, not 6-month post-program measure. A follow-up measure was taken 4 months later, which was 10 months post-baseline but only 4 months post-program, which is what IRE reported. IRE's focus was on the *post-program* duration of effects.
James, 2006 [2 data points]	Any condom use	Evidence of programme potential at 6 months. The students in the full implementation group reported significantly more condom use at last sex (B = −.80, SE = .40, Wald (1, 157) = 4.16, $p < .05$, OR = .45). The original report is misleading because the group that showed positive impact was the full implementation group. This was compared against the control group. This aligns with the original report's definition of programme potential because it was measured at 6 months, but it was still misreported that it reflected a subgroup.	Reported as evidence of programme potential: "at Prog End-Subgroup O".	1.This effect was reported at the post-test (i.e., at program end), not 6 months post-program (see above notes), for the "full implementation" group. 2.The "full implementation" group was not the full sample but actually a subgroup, since it was identified as "exploratory analyses" (p.287) that divided the full sample into those who received "full" or "partial" implementation of the program.[d] IRE took the "intent to treat" approach and considered these as subgroups, as the study authors also appeared to do, since they reported no overall behavioral effects for the full sample at any time. See statement of results in study abstract, p.281: "No effects were found on safe sex practices (condom use, sexual intercourse)."
James, 2006 [2 data points]	Frequent/recent sex	Evidence of programme potential at 6 months. The students in the full implementation group reported significantly less sexual activity in the previous 6 months (B = −.53, SE = .24, Wald (1, 657 = 4.98, $p < .05$, OR = .59). The original report is misleading because the group that showed positive impact was the full implementation group. This was compared against the control group. This aligns with the original report's definition of programme potential because it was measured at 6 months, but the original report misreported that it reflected a subgroup.	Reported as evidence of programme potential: "at Prog End-Subgroup O".	1. As explained in the above note (James, 2006), the "full implementation" group was a subgroup, not the full sample, since the analysis divided the sample into those who received "full" or "partial" implementation of the program. IRE took the "intent to treat" approach and considered these as subgroups, as the study authors also appeared to do, since they reported no overall behavioral effects *for the full sample* at any time.[d] This claim of IRE error, and the two immediately above it, are (like several others in this table) the result of a lack of close reading of the study text by the WHO reviewers, resulting in error on their part in interpreting the study data. 2. This outcome was measured at the post-test (end of program), not 6 months post-program. See above notes (james, 2006) on the 6-month post-baseline measurement.
Jemmott, 2015	Frequent/recent sex	Positive effect. Table 3 reports the odds ratios for vaginal intercourse in the past 3 months. The overall intervention effect was	Reported as NS.	The short-term outcome reported was an average of the program effects at the 3, 6, and 12-month follow-ups. This is not a measure of a 12-month post-program effect. Since

		not significant ($p = 0.076$), but the short term intervention effect (3, 6, 12 month) had an OR = 0.62 ($p = 0.022$). The insignificant overall follow up does not discount the significant data at 12 months.		program effects usually decline substantially over a 12-month time frame (they are often large at 3 mo. and nonexistent at 12 months), this averaging technique can produce a modest but statistically significant average effect over the entire time span, even though the program impact has actually disappeared after 12 months, both in quantity and statistical significance. This gives the appearance that the program's effect lasted 12 months after the program by making a non-significant effect at 12 months post-program look significant because the average "over the 12-month period" was significant. Thus, these "short-term effects" are not equivalent to measuring program effects after 12 months and shouldn't be viewed as such. For this reason IRE relied on the overall program effects, which were not significant.
Jemmott, 2015	# sex partners	Positive effect. Table 3 reports the odds ratios for multiple partners in the past 3 months. The overall intervention effect was not significant ($p = 0.095$), but the short term intervention effect (3, 6, 12 month) had an OR = 0.50 ($p = 0.0180$) (considered 12 months post-program).	Reported as NS.	The study authors (the Jemmotts) would prefer that the "short term" effects (the average of the 3, 6, and 12 month effects) be considered the same as 12-month post-program effects, but they are not. See the above comment for Jemmott, 2015, on the problem with the "short term effects" measurement.
Jemmott, 2015	STDs	Positive effect. The intervention group had significantly reduced curable STIs at 42-month follow-up, OR = 0.71, 95% CI [0.54, 0.95], but not at 54-month follow-up, OR = 1.15, 95% CI [0.84, 1.57]. The insignificant 54-month follow up does not discount the significant data at 54 months. These data reflect the entire intervention group, so stating subgroup is incorrect. Because the data are reported at 42 months, this aligns with the definition of positive effect.	Reported as evidence of programme potential at 42 months, noting "Subgroup O".	Per Table 2, the OR of .71 is not for the entire intervention group but for the subgroup who report themselves as sexually experienced. However, the sexually inexperienced intervention subgroup group showed 11% as testing positive for STDs, compared to only 4% in the control group. When the sexually experienced and inexperienced groups are added together, the OR is .90. The study reports that this "main effect on STI was nonsignificant" (see Study Abstract). Thus, as IRE reported, there was no STD effect for the full sample.[d]
Jewkes, 2008	STDs	Positive effect. Table 3 reported significantly lower incidence of HSV-2 in Stepping Stones than control intervention for both men and women (incident rate ratio: 0.67, $p = 0.036$). There was lower incidence of HIV in Stepping stones than control, but was not significant. These data were normalised across both follow-ups at 12 and 24 months.	Reported as positive and negative effect.	The outcome for STDs was not reported as both a positive and negative effect by IRE. This is a misunderstanding of the color-coding on the IRE data table (see above note in yellow for Dupas, 2011). The grey color did not mean there were both positive and negative effects on the STDs outcome, but that *the program* produced both positive and negative effects across the targeted outcomes. This was noted in the color key at the bottom of Table 7, but could have been stated more clearly. In Table 7, the STD outcome was recorded as a positive 12-month effect, but color-coded grey rather than green because it was not Evidence of Program Effectiveness, due to the occurrence of other, harmful program effects. The negative effect by this program was specified in Column 4 of Table 7 as being for Paid Sex, not STDs. This negative program effect disqualified the program for designation as an effective program, according to standards of the scientific field of program effectiveness.[e]
Jewkes, 2008	Frequent/recent sex	NM. Recent or frequent sex was not measured in this study. There was a negative impact for transactional sex, which is captured in the "Negative effect" column. Transactional sex is not recent sex. This is misleading and incorrect.	Reported as a negative effect.	There was an increase in the frequency of a type of sexual activity: Paid Sex. While it does not fit the typical definition of Frequency of Sex, it was close enough to allow IRE to list it in this pre-existing column, without creating a new column for a rarely measured outcome. The type of Frequent Sex was specified in Column 4 of Table 7 of the IRE report under

				Negative Effects as Paid Sex. IRE concedes that this is a little confusing but disagrees that it was an IRE error to record the outcome this way.
Jewkes, 2008	Dual benefit (12 mo.)	NM. Sexual initiation is not measured (correctly marked by report). Therefore, this should be reported as NM too.	Reported as NS.	IRE concurs.
Karnell, 2006	Sexual Initiation	NM. Table 2 measures this at the baseline. It is included in other tables to analyse data by those who had and hadn't had sex at the pretest, but it is never re-measured after the intervention.	Reported as NS.	IRE concurs.
Li, 2008	Frequent/recent sex	NS. Table 2 reports results of whether participants have engaged in sexual inter-course in the past 6 months but $p > 0.05$.	Reported as NM.	IRE concurs.
Li, 2008	Sexual Initiation	NM	Reported as NS.	IRE concurs.
Magnani, 2005	Overall comment	Spelled the last name wrong	Spelled "Magnini"	This was a "typo" and should not be counted on a list of research discrepancies.
Martinez-Donate, 2004	Any condom use	NM. Condom acquisition was measured but use was not.	Reported as NS.	On p.269 the study says "We asked students…the frequency (never, sometimes, or always) of condom use" and that the information from this question was used to create the outcome measure Unprotected Sex. We coded this outcome as Condom Use, because that is specifically what was being measured, even though the study used the surrogate label of Unprotected Sex. This was not an error but a deliberate reasonable decision. (The label Unprotected Sex is unclear; the way it is measured sometimes encompasses LARC[c] use and/or abstinence.)
Mathews, 2010	Overall comment	Spelled the last name wrong	Spelled "Matthews"	This was a "typo" and should not be counted on a list of research discrepancies.
Mathews, 2010 – Site 3	Sexual Initiation	Positive effect. Table 3 presents adjusted odds ratio for sexual debut during the 12–15 month study to be 0.65 ($p > 0.05$) showing that study participants in this group are significantly less likely to have initiated sexual activity. This was clearly a positive impact and was measured for the entire study population (not a subgroup).	Reported as evidence of programme potential and indicated for subgroup.	When males and females were analyzed separately (p.117, Supplementary Tables 3 & 4), the effect was found only for males and not females. Thus, it became clear that it was a subgroup effect.[d]
Merakou, 2006	Sexual Initiation	Evidence of programme potential at 8 months. Table 4 presents responses to statement "I never had sexual relationships". The intervention group showed a positive effect decreasing by 14.2% ($p = 0.001$) and the control group decreasing by 10.5% ($p = 0.064$) but it was not significant. This confirms that the intervention group had a significant positive impact on sexual initiation.	Reported as negative effect.	On p. 130, the study states: "the % of students who started sexual relations… was higher in the intervention group" and on p.131, "more students from the intervention group initiated sex." The greater *decrease* in the intervention than the control group in teens who had *never* had sex (cited by VanTreeck from Table 4) was a greater *increase* in sexual initiation. While this could literally be called a "positive" effect, since initiation rates went *up*, it was a negative/harmful program impact on teens, not a positive/desired one. (This is the way IRE consistently used the term "negative effect," to mean an undesired effect on teen behavior). This increase in sexual initiation caused by the CSE program was not "evidence of program potential." The VanTreeck claim of IRE error on this outcome is itself a very basic error by VanTreeck in interpreting the study results.
Merakou, 2006	Any condom use	Evidence of programme potential at 8 months. Table 4 shows responses to statement "During the last year I initiated	Reported as evidence of programme potential for 6 months post baseline.	On p. 129, the study states: "The questionnaire was completed…before and after 6 months of the HIV prevention intervention." This was a 6-month post-baseline measure. The

		the use of a condom with my partner." In the intervention group agreeance to this statment increased by 1.3% ($p = 0.017$) where as the control group's agreeance decreased by 5.4% but was not significant. The measurement was taken 8 months post baseline (Oct 1997 to May 1998).		end of October 1997 to the beginning of May 1998, is roughly a 6-month timespan.
Merakou, 2006	Negative Effect	No. See error for Merakou, 2006: sexual initiation.	Reported negative effect for sexual initiation.	See the first comments above, for Merakou, 2006, noting the basic error by VanTreeck et al. in interpreting the study results.
Merakou, 2006	Dual benefit (12 mo.)	Evidence of programme potential at 8 months. See sexual initiation and any condom use.	Reported as NM.	These effects were not measured at 12 months post-program, which is the definition of IRE's Dual Benefit outcome measure: Per p. 129, "The questionnaire was completed…before and after 6 months of the HIV prevention intervention." This was a 6-month post-baseline measure.
Okonofua, 2003 [2 data points]	Any condom use	Evidence of programme potential at 11 months. Table 4 reports the odds ratios for condom use. The odds ratio for increased condom use pre- to post-intervention within study groups between both sexes was OR = 1.48 (95% CI = 1.22–1.79) in the intervention group and OR = l.ll (95% CI = O.86–1.42) in the two control groups combined.	Reported as evidence of programme potential at 1 year for a subgroup.	1.The comparison of interest is not the pre-post change within groups but the comparison of the intervention group relative to the control group. 2.The program effect is stated on p.66: "this statistically significant *effect of the intervention* was due to the reported increase among females (OR=1.80, 95%=1.11-2.92) rather than among males (OR=1.13, 95% CI=O.84-1.51) when *comparing the intervention to control groups*" (emphasis added). Thus, it was a subgroup effect on Condom Use, occurring for females only.[d]
Okonofua, 2004 [2 data points]	STDs	Evidence of programme potential at 11 months. Table 6 reports intervention vs. both controls both genders adjusted change for STD symptoms as 0.68 (significant at 95% CI) meaning that intervention participants were 0.68 times as likely as control participants to report STD symptoms.	Reported as positive effect at 1 year.	1. The study states in 2 places that the follow-up survey was conducted 1 year post-baseline, on p. 63, "the postintervention evaluation to be undertaken 1 year later" and p.71: "The study was designed to minimize loss to follow-up by conducting the pre- and post-intervention surveys among adolescents in senior classes 4 and 5, to ensure their availability 12 months after baseline." 2. The year for this study was 2003, not 2004.
Ross, 2007	# sex partners	Evidence of programme potential for a subgroup at 3 years. Significant effect in males but not females. Table 4 shows RR with 95% CI for reporting more than 1 partner in past 12 months. Males: 0.69 (0.49, 0.95), positive intervention effect. Females: 1.04 (0.58, 1.89), not significant. Data demonstrate positive effect and insignificant data.	Reported as positive and negative effect.	It is Ross' Table 3, not Table 4 showing these results. Per Table 3, IRE reported a positive subgroup effect at 3 years post-baseline for Number of Sex Partners and a negative effect for STDs (gonorrhea for females: ARR=1.93 (CI=1.01-3.71)). This shows as a negative effect on STDs in Column 4 of IRE Table 7. Number of Sex Partners is not listed under negative effects in Column 4. Thus, IRE did not report both a positive and negative effect for Number of Sex Partners. This is a misunderstanding of the color-coding on the IRE data table (see above note in yellow for Dupas, 2011). The grey color did not mean there were both positive and negative effects on that specific outcome, but that *the program* produced both positive and negative effects across the targeted outcomes. This was noted in the color key at the bottom of Table 7, but could have been stated more clearly. The negative program effect on STDs disqualifies this CSE program for designation as showing "program potential," according to standards of the scientific field of program effectiveness.[e]

Ross, 2007 [2 data points]	Any condom use	Evidence of programme potential for a subgroup at 3 years. Significant effect in males but not females. Table 3 shows RR with 95% CI for used condom at last sex. Males: 1.47 (1.12, 1.93), positive intervention effect. Females: 1.12 (0.85, 1.48), not significant. Data demonstrate positive effect and insignificant data.	Reported as positive and negative effect.	1. There were 2 measures of condom use: "initiation of" and "use at last sex." There were significant effects for young men and young women on the former, so IRE gave the program the benefit of the doubt and attributed an effect on condom use. This was a reasonable decision, not an error. 2. This is a misunderstanding of the color-coding on the IRE data table (see above notes in yellow for Dupas, 2011, and Ross, 2007). IRE did not report a negative effect for condom use in the Negative Effects column (Column 4 of Table 7), but for STDs. The negative effect on STDs disqualifies this CSE program for designation as showing "program potential," according to the standards of the scientific field of program effectiveness.[e]
Shuey, 1999	Dual benefit (12 mo.)	NM. Condom use is NM and was reported correctly, therefore dual benefit should be reported as NM.	Reported as NS.	IRE concurs.
Stanton, 1998	Sexual Initiation	Positive effect. Table 1 shows that the number of remaining virgins (among those who were virgins at baseline) in intervention group was 17% ($p < 0.05$) and the number of remaining virgins in the control group was 9% (not significant) 12 months post-intervention. This demonstrates that the intervention had a significant positive effect on sexual initiation, whereas the control group did not. The measurement of remaining virgins was taken for the entire study population (not a subgroup).	Reported as evidence of programme potential at 12 months for a subgroup.	On p.2475, the study states: "This [program] effect appears to have been contributed primarily by changes among the females (e.g. there was no statistically significant effect on abstinence among males who were virgins at baseline)." Thus, it was actually a subgroup effect—occurring for females but not for males, according to the study authors.[d]
Stanton, 1998	Any condom use	Evidence of programme potential for a subgroup immediately post-intervention. Baseline virgins. Table 3 shows that the 6 and 12 mo f/u are not significant.	Reported as evidence of programme potential at 6 months for a subgroup.	IRE concurs.
Stephenson, 2008	STDs	NS. Table 2 and 3 measure ever been told by a doctor or nurse that you have an STD but results were NS.	Reported as NM.	IRE concurs.
Taylor, 2014	Any condom use	Evidence of programme potential at 8 months. Table 2 reports the intervention effect on condom use to be positive; 0.98 ($p < 0.01$).	Reported as evidence of programme potential at 5 months.	The program duration was 3 months (p.849), so the 8-month post-baseline follow-up measure was actually a 5-month post-program effect, as correctly recorded by IRE. The focus of the IRE analysis was on post-program effects.
Taylor, 2014	Dual benefit (12 mo.)	NS. Condom use and sexual initiation were both NS.	Reported as NM.	These effects were not measured at 12 months post program (see above comment, Taylor, 2014), which is required for IRE's Dual Benefit outcome; thus, a 12-month Dual Benefit was not measured.
Visser, 2007	Sexual Initiation	Positive effect. Table 3 reports that the intervention group sexual experience remained the same 18-months post-intervention (no significant change). The control group sexual experience increased significantly ($p < 0.05$). This demonstrates a positive intervention effect because the intervention significantly reduced sexual initiation.	Reported as positive and negative effect.	This is a misunderstanding of the color-coding on the IRE data table (see above note in yellow for Dupas, 2011). IRE reported only a positive effect for Sexual Initiation. The Negative Effects are clearly labeled as Forced Sex and Number of Sex Partners, in Column 4 of Table 7 in IRE's report. Despite its positive effect on Sexual Initiation, this program does not qualify as showing evidence of effectiveness because of its multiple negative/ harmful program impacts, according to standards of the scientific field of program effectiveness.[e]
Visser, 2007	Frequent/recent sex	Evidence of programme potential at 18 months. Table 3 presents data for having had sex in the past 3 months. The intervention	Reported as positive and negative effect.	This is a misunderstanding of the color-coding on the IRE data table (see above note in yellow for Dupas, 2011). IRE reported only a positive effect for Frequent/Recent Sex. The Negative

		group did not see a significant change, but the control group saw a significant increase ($p < 0.001$) in recent sex.		Effects are clearly labeled as Forced Sex and Number of Sex Partners, in Column 4 of Table 7 in IRE's report. According to scientific standards of program effectiveness, this CSE program does not qualify as showing evidence of program potential because of its multiple negative/harmful program impacts.[e]
Walker, 2006	Frequent/recent sex	NS. Table 1 compares participants in each group who have been sexually active. It is not explicitly defined to be initiation, so it is assumed it is for recent sex.	Reported as NM.	Table 1 is the baseline characteristics of the sample at the 16-month follow-up, not a report of program effects. Tables 2 & 3 are program effects for the sexually active portion of the sample (p.3), with no outcomes reported for sexual activity itself. So Frequent/Recent Sex was not measured/reported as a program outcome.
Wight, 2002	Dual benefit (12 mo.)	NS. Any condom use and sexual initiation were both NS.	Reported as NM.	P.2 reports that the study "recruited two successive cohorts of third year secondary school pupils (aged 13-14 years) in 1996 and 1997 and followed them up at the start of their fifth year (at 15-16 years), about six months after completion of the programme." Thus, the follow-up was only 6 months post-program; a 12-month post-program Dual Benefit was not measured.
Ye, 2009	Any condom use	Evidence of programme potential at 1 month. The study measured condom use if intercourse happened. At one month follow-up there was a significant increase of condom use in the intervention group ($p < 0.01$). There was no significant effect between one month and one year follow-up.	Reported as NS.	IRE's information came from the abstract of this study and from data about it provided in the Fonner, 2014 meta-analysis.[f] The full text of the Ye, 2009, study was not available to IRE. The study abstract did not report any program impact on condom use or on any behavior at any follow-up point. The Fonner study indicated that Condom Use was measured but no program impact was reported. The Ye abstract reports the long-term follow-up time as 2 years, not 1 year as reported by VanTreeck. We do not consider effects at 1-month post-program as evidence of program potential. For these reasons, we do not consider IRE's reporting of the data to be in error.

Key to Color Code:

VIOLET = A mistaken interpretation of this particular study's findings by the WHO reviewers, not a discrepancy between the IRE report and the study findings.
YELLOW = A misunderstanding of the IRE Table 7 notations by the WHO reviewers, not a discrepancy between the IRE report and study findings.
BLUE = A reasonable decision by IRE on interpretation of this study's findings, not a discrepancy between IRE and the study findings.
GREEN = IRE concurs with WHO reviewers that there is a discrepancy between the IRE report and this particular study's findings (only on the outcome indicated).

Summary of IRE Analysis of Purported Discrepancies:

In Table B1, VanTreeck, et al., claim there are 66 discrepancies between the IRE findings in Weed and Ericksen, 2019 (Table 7), and the findings of the 43 studies they reviewed. There are 59 discrepancy entries in Table B1, with 7 of these naming 2 different discrepancies within the single entry, for a total of 66 data points. IRE re-examined the results of each of the pertinent studies and found 9 of the 66 items were verified as actual discrepancies with the study results. There were 430 data points in IRE's Table 7 (43 reviewed studies x 10 potential outcomes for each). This gave an error rate of 9 discrepancies out of 430 data points, or 2%. The remaining 57 proposed discrepancies were not confirmed as discrepancies between the IRE report and the 43 reviewed studies. Twelve were reasonable decisions by IRE in the interpretation of study results, 8 were misunderstandings by the WHO reviewers of the notations in IRE's Table 7, and 37 were erroneous interpretations of study results by VanTreeck, et al. (The justifications for each of these designations are documented in detail above, in the column to the right of Table B1.) The 37 mistaken interpretations by WHO reviewers yielded an error rate of 56% (37 ÷ 66). The 9 confirmed IRE discrepancies are all minor in nature and do not change the conclusions of the IRE analysis regarding numbers of CSE programs showing Evidence of Effectiveness (3) or Negative Effects (9).

Endnotes

a. Table B1 is presented here as it is shown in the VanTreeck, et al., 2023 article (in References, below), with the exception of a few notes in brackets that were added by IRE for clarification. The purported discrepancies are based on the data in Table 7 of the IRE report: Weed SE, Ericksen IH. (2019). Re-Examining the Evidence for Comprehensive Sex Education in Schools: A Global Research Review. See: *https://www.institute-research.com/CSEReport/Global_CSE_Report_12-17-19.pdf;* see also, Ericksen, I.H. and Weed, S.E. (2019). "Re-Examining the Evidence for School-based Comprehensive Sex Education: A Global Research Review." *Issues in Law and Medicine, 34(2):161-182.*
b. The content entered in these columns is an interpretation by the WHO reviewers of the results of the designated study and the findings of the IRE report. Their interpretation is not always an accurate report of the study results or the IRE findings. These discrepancies are explained in the colored sections in the righthand column.
c. LARC stands for Long-Acting Reversible Contraception method, such as the non-hormonal copper coil (IUD), hormonal coil (IUS), and Depo-Provera injection.
d. *The Society for Prevention Research*, in its "Standards for Effectiveness" says, "The degree to which findings are generalizable should be evaluated…Statistical analysis of subgroup effects must be conducted for each important subgroup to which intervention effects are generalized [to enable the]…demonstration of robust effects across important population subgroups" and thereby detect if program effects are generalizable across the target population of the intervention. See: Gottfredson DC, Cook TD, Gardner FEM, Gorman-Smith D, Howe GW, Sandler IN, Zafft KM. (2015). Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research in Prevention Science: Next Generation. *Prev Sci*, 16(7):893-926, p. 18-19.
e. *The Society for Prevention Research* stipulates, regarding prevention programs: "For an efficacy claim, there must be no serious negative (iatrogenic) effects on important outcomes." See: Gottfredson DC, Cook TD, Gardner FEM, Gorman-Smith D, Howe GW, Sandler IN, Zafft KM. (2015). Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research in Prevention Science: Next Generation. *Prev Sci*, 16(7):893-926, p.17, Standard 7c. This standard is also held by *Blueprints for Healthy Youth Development*: Blueprints Standards. Available at: https://www.blueprintsprograms.org/blueprints-standards/
f. Fonner VA, Armstrong KS, Kennedy CE, O'Reilly KR, Sweat MD. (2014). School Based Sex Education and HIV Prevention in Low and Middle-Income Countries: A Systematic Review and Meta-Analysis. Plos One (9), e89692.

References

Aderibigbe SA, Araoye MO. (2008). Effect of Health Education on Sexual Behaviour of Students of Public Secondary Schools in Ilorin, Nigeria. *European Journal of Scientific Research*, 24 (1) 33-41.

Agha S, Van Rossem R. (2004). Impact of a School-based Peer Sexual Health Intervention on Normative Beliefs, Risk Perceptions, and Sexual Behavior of Zambian Adolescents. *JAdolescHealth*; 34:441–452.

Ajuwon A, Brieger W. (2007). Evaluation of a school-based Reproductive Health Education Program in rural South Western, Nigeria. *African Journal of Reproductive Health;* 11 (2), 47-59.

Borgia P, Marinacci C, Schifano P, Perucci C. (2005). Is peer education the best approach for HIV prevention in schools? Findings from a randomized controlled trial. *JAdolescHealth*; 36, 508–516.

Dente, M, Fabiani, M, Okwey, R, Conestà, N, et al. (2005). Impact of Voluntary Counselling and Testing and Health Education on HIV Prevention among Secondary School Students in Northern Uganda. *VCT AND HEALTH EDUCATION FOR HIV PREVENTION*; 3 (1) 1 – 11

Díaz M, de Mello MB, de Sousa MH, Cabral F, de Castro e Silva R, Márcia Campos M, Faúndes A. (2005). Outcomes of three different models for sex education and citizenship programs concerning knowledge, attitudes, and behavior of Brazilian adolescents. *Cad. Saúde Pública, Rio de Janeiro*, 21(2):589-597, mar-abr, 2005.

Doyle AM, Ross DA, Maganja K, Baisley K, Masesa C, Andreasen A, Plummer ML, et al. (2010). Long-Term Biological and Behavioural Impact of an Adolescent Sexual Health Intervention in Tanzania: Follow-up Survey of the Community-Based MEMA kwa Vijana Trial. *PLoS Medicine*, 7 (6), e1000287

Duflo E, Dupas P, Kremer M, Sinei S. (2006). Education and HIV/AIDS Prevention: Evidence from a randomized evaluation in Western Kenya. *World Bank Policy Research Working Paper 4024*, October 2006.

Duflo E, Dupas P, Kremer M. (2015). Education, HIV, and Early Fertility: Experimental Evidence from Kenya. *Am Econ Rev*. 2015 September ; 105(9): 2757–2797.

Dupas P. (2011). Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya. *American Economic Journal: Applied Economics 3: 1–34.*

Fitzgerald AM, Stanton BF, Terreri N, Shipena H, Li X, et al. (1999). Use of Western-Based HIV Risk-Reduction Interventions Targeting Adolescents in an African Setting. *JAdolescHealth;* 25:52–61.

James S, Reddy P, Ruiter R, McCauley A, van den Borne B. (2006). The impact of an HIV and AIDS life skills program on secondary school students in Kwazulu-Natal, South Africa. *AIDS Education and Prevention*, 18(4), 281–294 Jemmott, J, Jemmott, L, Oleary, A, Ngwane, Z, et al. (2015). HIV/STI Risk-Reduction- Intervention Efficacy with South African Adolescents Over 54 Months. *Health Psychology*: 34(6): 610–621

Jemmott JB, Jemmott LS, O'Leary A, Ngwane Z, Lewis DA, et al. (2015). HIV/STI Risk-Reduction-Intervention Efficacy with South African Adolescents Over 54 Months. *Health Psychol*; 34(6): 610–621.

Jewkes R, Nduna M, Levin J, Jama N, et al. (2008). Impact of Stepping Stones on incidence of HIV and HSV-2 and sexual behaviour in rural South Africa: cluster randomised controlled trial. *BMJ*; 337:a506 doi:10.1136/bmj.a506

Karnell AP, Cupp PK, Zimmerman RS, Feist-Price S, Bennie T. (2006). Efficacy of an American alcohol and HIV prevention curriculum adapted for use in South Africa: results of a pilot study in five township schools. *AIDS Education and Prevention*, 18(4), 295–310.

Li X, Stanton B, Wang B, Mao R, Zhang H, et al. (2008). Cultural adaptation of the Focus on Kids program for college students in China. *AIDS Education and Prevention*, 20(1), 1–14.

Magnani R, MacIntyre K, Karim AM, Brown L, Hutchinson P. (2005). The impact of life skills education on adolescent sexual risk behaviors in KwaZulu-Natal, South Africa. *JAdolescHealth*; 36, 289–304

Martinez-Donate AP, Hovell MR, Zellner J, Sipan CL, Blumberg EJ, Carrizosa C. (2004). Evaluation of Two School-Based HIV Prevention Interventions in the Border City of Tijuana, Mexico. *The Journal of Sex Research,* 41 (3) 267-278.

Mathews C, Aarø LE, Grimsrud A, Flisher AJ, et al. (2012—listed as 2010 in IRE report). Effects of the SATZ teacher-led school HIV prevention programmes on adolescent sexual behaviour: cluster randomised controlled trials in three sub-Saharan African sites. *International Health*, (4) 111– 122, Site 3.

Merakou K, Kourea-Kremastinou J. (2006). Peer education in HIV prevention: an evaluation in schools. *European Journal of Public Health*, Vol. 16, No. 2, 128–132. (

Okonofua FE, Coplan P, Collins S, Oronsaye F, et al. (2003). Impact of an intervention to improve treatment-seeking behavior and prevent sexually transmitted diseases among Nigerian youths. *Int J Infect Dis*; 7: 61-73

Ross DA, Changaluchab J, Obasia AIN, Todd J, Plummera ML, et al. (2007). Biological and behavioural impact of an adolescent sexual health intervention in Tanzania: a community-randomized trial. *AIDS*, 21:1943–1955

Shuey DA, Babishangire BB, Omiat S, and Bagarukayo H. (1999). Increased sexual abstinence among in-school adolescents as a result of school health education in Soroti district, Uganda. *HEALTH EDUCATION RESEARCH*, 14 (3), 411–419

Stanton BF, Li X, Kahihuata J, Fitzgerald A, et al. (1998). Increased protected sex and abstinence among Namibian youth following a HIV risk-reduction intervention: a randomized, longitudinal study. *AIDS* 1998, 12:2473–2480

Stephenson J, Strange V, Allen E, Copas A, Johnson A, et al. (2008). The long-term effects of a peer-led sex education programme (RIPPLE): A cluster randomised trial in schools in England. *PLoS Med,* 5(11): e224.

Taylor M, Jinabhai C, Dlamini S, Sathiparsad R, Eggers MS, De Vries H. (2014). Effects of a Teenage Pregnancy Prevention Program in KwaZulu-Natal, South Africa, *Health Care for Women International*, 35:7-9, 845-858.

VanTreeck K, Elnakib S, & Chandra-Mouli V. (2023) A reanalysis of the Institute for Research and Evaluation report that challenges non-US, school-based comprehensive sexuality education evidence base. *Sexual and Reproductive Health Matters*, 31:1, 2237791, DOI: 10.1080/26410397.2023.2237791

Visser M. (2007). HIV/AIDS prevention through peer education and support in secondary schools in South Africa, SAHARA-J: *Journal of Social Aspects of HIV/AIDS*,4:3, 678-694, DOI: 10.1080/17290376.2007.9724891

Weed SE, Ericksen IH. (2019). Re-Examining the Evidence for Comprehensive Sex Education in Schools: A Global Research Review. *https://www.institute-research.com/CSEReport/Global_CSE_Report_12-17-19.pdf*

Walker D, Gutierrez JP, Torres P, Bertozzi SM. (2006). HIV prevention in Mexican schools: prospective randomized evaluation of intervention, BMJ, doi:10.1136/bmj.38796.457407.80

Wight D, Raab GM, Henderson M, Abraham C, Buston K, et al. (2002). Limits of teacher delivered sex education: interim behavioural outcomes from randomised trial. *BMJ*; 324, 15 JUNE 2002, bmj.com

Ye XX, Huang H, Li SH, Xu G, Cai Y, et al. (2009) HIV/AIDS education effects on behaviour among senior high school students in a medium-sized city in China. *Int J STD AIDS;* 20: 549–552.